Effect of Bounded Rationality on Tradeoff Between Systemic Risks & Economic Efficiency in Networks

V. Marbukh
National Institute of Standards and Technology
100 Bureau Drive, Stop 8910
Gaithersburg, MD 20899-8910
E-mail: marbukh@nist.gov

Abstract—**This paper discusses effect of bounded rationality on the systemic risks vs. economic performance tradeoff in large-scale networks operating under economic pressures. Existence of systemic risks in economically incentivized networked systems is demonstrated by recent numerous systemic failures in various critical large-scale networked infrastructures. Using "Complex System" perspective, we consider systemic risks due to overload experienced by a sizable portion of the network. Economic pressures facilitate systemic overload by incentivizing (a) high level of dynamic resource sharing allowing the system to mitigate effects of inherently uncertain exogenous demand and limited system reliability, and (b) network operation on the stability boundary where all network resources are fully utilized. However, in practice these economic incentives are counteracted by bounded rationality of the network operator(s), e.g., due to limited information on the uncertain exogenous environment. We argue that bounded rationality reduces both the network operational region and risk of abrupt overload. Due to higher performance losses and lower predictability of a discontinuous/abrupt overload as compared to a continuous/gradual one, our results suggest that bounded rationality may benefit the system performance by allowing system to operate closer to the stability boundary.**

Keywords-large scale network, economic efficiency, systemic risk, bounded rationality.

I. INTRODUCTION

Economic and convenience benefits of interconnectivity drive current explosive emergence and growth of networked systems [1]. One of these benefits is ability of interconnected systems to support dynamic resource sharing allowing for sustaining certain level of resource demand/supply imbalances due to exogenous demand variability and limited reliability of system component. This increase in the network "robustness" can be quantified by the corresponding enlargement of the network operational region. However, numerous recent systemic failures in various "performance-oriented" networked infrastructures demonstrated existence of systemic risks associated with economic benefits of interconnectivity. These systemic risks may be a result of the same economic incentives driving networked systems toward the stability boundary, where system resources are fully utilized.

Our previous results [2]-[4] suggested that inherent tension between trends for (a) enlargement of the network operational region on the one hand and (b) keeping system close to the boundary of this region on the other hand is a form of "robust yet fragile" phenomenon [5]. Here robustness to "sufficiently" small resource supply/demand imbalances is due to enlargement of the operational region, while fragility is due to increased risk of discontinuous/abrupt systemic instability as the boundary of the operational region is breached in a case of "sufficiently" large resource demand/supply imbalances. This interpretation is based on higher performance losses and lower predictability associated with abrupt/discontinuous as compared to gradual/continuous instabilities.

This paper argues that bounded rationality may allow system operator(s) to manage this robustness/fragility tradeoff by reducing risk of abrupt instability at the cost of reduction of the operational region. Thus bounded rationality may benefit the system performance by allowing system to operate closer to the stability boundary. Despite our assessments are based on analysis of a homogeneous network, our previous results on Perron-Frobenius characterization of systemic instabilities [2]-[4] indicate applicability of these assessments to heterogeneous networks. The paper is organized as follows. Section II introduces model of a networked system of shared resources. The system efficiency is controlled through pricing of the exogenous elastic demand, where bounded rationality of the controller(s) is due to uncertainties in the price-demand curve. Section III introduces a mean-field and fluid approximate performance models, and discusses effect of bounded rationality on the system performance.

II. NETWORK: BOUNDED RATIONALITY

Following [6] consider the following model of Cloud computing system [7]. The system includes I classes of jobs (requests) and J service groups, where group $j=1,..,J$ includes N_j servers and a buffer capable of holding up to B_j jobs. Jobs of class $i=1,..,I$ arrive following a Poisson process of rate Λ_i. Different service groups may include geographically distant resources and different types of resources, e.g., memory, CPUs, communication resources, etc. A job of class i can be serviced on one of several resource sets. Service times are distributed exponentially with service group specific averages. We assume a service strategy which either rejects or accepts an arriving job. In the latter case the job stays until service is completed. We also assume a work-conserving

service discipline which does not allow an idle server in a group with at least one buffered job.

An arriving class $i=1,..,I$ job has exponentially distributed service time with average $1/\mu_{ij}$ on a class $j=1,..,I$ server. Static routing strategy is characterized by probabilities q_{ij} that an arriving request of class i is routed to server group j, where rejection probabilities $q_{i0} := 1-\sum_j q_{ij}$ characterize admission strategy. Economically incentivized network operator attempts to maximize the generated revenue rate. In particular, assuming that rates $\Lambda_i = \Lambda_i(p_i)$ are decreasing functions of price p_i, which provider charges a class i request for service, and moreover demand is elastic [8], rational network operator attempts to maximize the aggregate revenue rate

$$R(p) = \sum_i p_i \Lambda_i(p_i) \left(\sum_j q_{ij}(1-\pi_j) \right) \quad (1)$$

over the price vector $p=(p_i)$. It can be shown [9] that at this optimum, exogenous demand and system capacity are matched, i.e., system attempts to accommodate the entire demand: $\sum_j q_{ij} = 1,\ i=1,..,I$ and system has no spare capacity: $\rho_j := (1/N_j)\sum_i q_{ij}\Lambda_i/\mu_{ij} = 1,\ j=1,..,J$.

We assume that the revenue maximization is subject to bounded rationality resulted from system operator(s) inability to completely control the exogenous demand through pricing, e.g., due to uncertainty in the price-demand curve $\Lambda_i(p)$. We model this uncertainty by assuming that rates Λ_i are random variables with averages $\widetilde{\Lambda}_i$ and standard deviations $\sigma(\Lambda_i)>0$. Network operator(s) being unable to control rates Λ_i, can control averages $\widetilde{\Lambda}_i = \widetilde{\Lambda}_i(p_i)$ by manipulating prices p_i.

In particular, often assumed form price-demand curve is [9]-[10] $\Lambda_i(p) = \Lambda_{i0} p^{-\varepsilon_i}$, where $\Lambda_{i0}>0$ and $\varepsilon_i > 1$ are the demand potential and elasticity respectively. Due to low reliability of numerical estimates of the demand potentials Λ_{i0} it is natural to assume that Λ_{i0} are random variables with averages $\widetilde{\Lambda}_{i0}$ and standard deviations $\sigma(\Lambda_{i0})>0$. In this model of bounded rationality, network operator(s) are only capable of controlling the expected exogenous demand $\widetilde{\Lambda}_i(p_i) = \widetilde{\Lambda}_{i0} p_i^{-\varepsilon_i}$ with standard deviation $\sigma(\Lambda_i(p_i)) = \sigma(\Lambda_{i0}) p_i^{-\varepsilon_i}$, by manipulating prices p_i.

III. Effect of Bounded Rationality

Introduce vector $\delta = (\delta_j, j \in J)$, where $\delta_j = 0$ if server group j has available resources, i.e., a server, or buffering space, or both. Otherwise $\delta_j = 1$. Since according to our assumptions $\bar{\delta}_j := E[\delta_j] << 1$, $j=1,..,J$, the main effect of dynamic resource sharing can be described by conditional rerouting probabilities q_{ijk} that a class i request initially routed to server group j is immediately rerouted to server group k in an unlikely case $\delta_j = 1$.

Assuming that exogenous demand has been already optimized over pricing, the system performance is characterized by loss probabilities. For a dynamic resource sharing discipline allowing a single rerouting attempt with probabilities q_{ijk}, loss probability for an arriving request of class i is

$$L_i = \sum_j q_{ij} \bar{\delta}_j \sum_{k \neq j} q_{ijk} E[\delta_k | \delta_j = 1], \quad (2)$$

where $\bar{\delta}_j := E[\delta_j]$ is the unconditional expectations of δ_j, and $E[\delta_k | \delta_j = 1]$ is the expectation of δ_k, given $\delta_j = 1$.

Our analysis is based on a mean-field type approximation [2]-[4], which neglects correlations between blockings in different service groups: $E[\delta_k | \delta_j = 1] \approx E[\delta_k]$, and thus allows us to approximate loss (2) as follows:

$$L_i \approx \widetilde{L}_i = \sum_j q_{ij} \bar{\delta}_j \sum_{k \neq j} q_{ijk} \bar{\delta}_k, \quad (3)$$

Dynamic resource sharing results in additional load due to allowing requests a second attempt to obtain service. The corresponding additional utilization for server group j is

$$\beta_j = \frac{1}{N_j} \sum_i \frac{\Lambda_i}{\mu_{ij}} \sum_{k \neq j} q_{ik} q_{ikj} \bar{\delta}_k, \quad (4)$$

We propose to approximate probabilities $\bar{\delta}_j$ by the Erlang formula with S_j servers, B_j, and utilization $\rho_j + \beta_j$ [2]-[4]:

$$\bar{\delta}_j \approx Erl(\rho_j + \beta_j, S_j, B_j) \quad (5)$$

where

$$Erl(\rho, S, B) = \frac{\frac{(S\rho)^{S+B}}{S! S^B}}{\sum_{k=0}^{S} \frac{(S\rho)^k}{k!} + \frac{(S\rho)^{S+B}}{S!} \frac{1-\rho^{B+1}}{1-\rho}} \quad (6)$$

Equations (4)-(6) form a closed mean-field approximation. Fluid approximation describes a limit of large service groups $S_j + B_j >> 1$, when equation (5) takes the following form [6]:

$$\tilde{\delta}_j = [1 - 1/(\rho_j + \beta_j)]^+, \quad (7)$$

since $(1-1/\rho)^+ = \lim_{S+B\to\infty} Erl(\rho, S, B)$, where $[x]^+ := \max(0, x)$. Combining (7) with (4) we obtain a closed system of fixed-point equations for $\tilde{\delta}_j$, $j = 1,..,J$.

Consider a particular case of symmetric system with "native services", where $N_i = N$, $\Lambda_i = \Lambda$, $\mu_{ii} = \mu$, $\mu_{ij} = \mu/(1+\chi)$; $i, j = 1,..,I$, $i \neq j$, parameter $\chi \geq 0$ characterizes inefficiency of a non-native service as compared to the native service, and dynamic resource sharing is characterized by probabilities $q_{ijk} = q$, $k \neq i, j$. In this case $\beta = (1+\chi)q\rho\bar{\delta}$, and thus (7) takes form:

$$\tilde{\delta} = \left[1 - \frac{1}{[1+(1+\chi)q\tilde{\delta}]\rho}\right]^+, \quad (8)$$

where exogenous utilization is $\rho := \Lambda/(N\mu)$.

Figures 1a and 1b sketch loss $\tilde{L}$ in a case $(1+\chi)q \leq 1$ of "soft" stability loss and in a case $(1+\chi)q > 1$ of "hard" stability loss respectively.

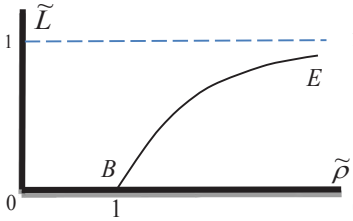

Figure 1a. $(1+\chi)q \leq 1$

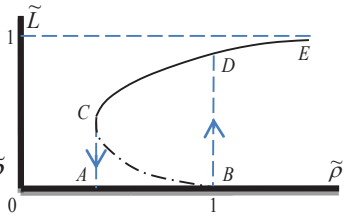

Figure 1b. $(1+\chi)q > 1$

Bounded rationality is modeled by assuming that utilization ρ is a random variable with average $\tilde{\rho}$ and standard deviation σ, and thus equation (8) is replaced with the following equation:

$$\tilde{\delta} = E_\rho\left\{\left[1 - \frac{1}{[1+(1+\chi)q\tilde{\delta}]\rho}\right]^+\right\} \quad (9)$$

Figure 2 sketches phase diagram of equation (9) in variables $(\tilde{\rho}, \sigma)$.

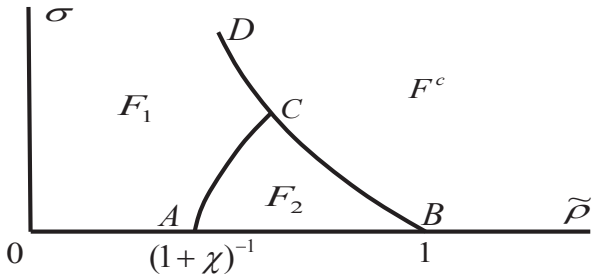

Figure 2. Phase diagram of equation (9).

In region $F := F_1 \bigcup F_2$ operational equilibrium $\tilde{\delta} = 0$ of equation (9) is asymptotically stable. In region F_1 equilibrium $\tilde{\delta} = 0$ is globally stable, while in region F_2 this equilibrium is locally stable and coexists with "undesirable" locally stable equilibrium $\tilde{\delta}^* > 0$. Following accepted practice, we associate multiple locally stable equilibria with metastable, i.e., persistent system states. Thus region F_1 represents "safe" operational region since breaching boundary DC results in "soft" stability loss similar to shown in Figure 1a.

Lower values of σ correspond to higher level of rationality resulting in widening operational region with respect to $\tilde{\rho}$, and thus to the expected serviced exogenous demand. However, this system robustness to variability of ρ within F_2 comes at the costs of operational equilibrium $\tilde{\delta} = 0$ fragility due to metastability of this equilibrium and "hard" stability loss as stability boundary BC is breached, i.e., due to "sufficiently large" variability in the instantaneous or long-term exogenous utilization. Figure 2 indicates that bounded rationality while reducing the stability region may also enlarge "safe" stability region with respect to $\tilde{\rho}$, and thus may benefit the system performance. This phenomenon is consistent with results [11]. Finally note that despite we analyzed a homogeneous network, our previous results on Perron-Frobenius characterization of systemic instabilities [2]-[4] indicate that our assessments can be extended to heterogeneous networks.

REFERENCES

[1] D. Helbing, Globally networked risks and how to respond, *Nature*. 497, 51-59, (02 May 2013).

[2] V. Marbukh, "Towards unified Perron-Frobenius framework for managing systemic risk in networked systems," European Safety and Reliability Conference (ESREL 2015), Zurich, Switzerland, 2015

[3] V. Marbukh, "On systemic risk in the cloud computing model," 26th International Teletraffic Congress (ITC), Karlskrona, Sweden, 2014.

[4] V. Marbukh, "Perron-Frobenius measure of systemic risk of cascading overload in complex clouds: work in progress," *IFIP/IEEE International Symposium on Integrated Network Management*, Gent, Belgium, 2013.

[5] J.C. Doyle, D.L. Alderson, L.Li, Steven Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger, "The ''robust yet fragile'' nature of the Internet," PNAS, vol. 102, no. 41 October 11, 2005, pp. 14497–14502.

[6] A.L. Stolyar and E.Yudovina, Systems with large flexible server pools: Instability of "natural load balancing," Annals of Applied Probability, Vol. 23, No. 5, 2013, pp. 2099-2138.

[7] The NIST Definition of Cloud Computing NIST Special Publication 800-145, http://csrc.nist.gov/publications/PubsSPs.html#800-145.

[8] D. Mitra, and Q. Wang, "Stochastic traffic engineering for demand uncertainty and risk-aware network revenue management," IEEE/ACM Trans.. on Networking, Vol. 13, No. 2, 2005, pp. 221-233.

[9] D. Xu, Y. Li, M. Chiang, and A. Calderbank, "Elastic service availability: utility framework and optimal provisioning," JSAC, 26, 6, 2008, 55-65.

[10] V. Marbukh and K. Mills, "Demand pricing & resource allocation in market based compute grids: a model and initial results," ICN 2008.

[11] G. Theodorakopoulos, J.Y. Le Boudec, and J. S. Baras, Selfish re-sponse to epidemic propagation. IEEE Trans. Aut. Contr., 58(2):363–376, 2013.